

—SISTEMAS DE RECOMENDACIONES—

UN SISTEMA DE RECOMENDACIONES TIENE COMO OBJETIVO DESCUBRIR CONTENIDOS QUE PUEDAN SER INTERESANTES PARA EL USUARIO. ESTOS SISTEMAS DESCUBREN CONTENIDOS QUE PUEDEN SER DE INTERESES PARA EL USUARIO Y SE LOS PRESENTA, EN LUGAR DE HACER QUE EL USUARIO BUSQUE HACERLO QUE LLEGUEN AL USUARIO. ENTONCES, EL USUARIO ENCUENTRA COSAS QUE NO SABÍA QUE EXISTIAN.

CARACTERÍSTICAS:

UN SISTEMA DE RECOMENDACIÓN EXITOSO DEBE CUMPLIR CON ESTAS CARACTERÍSTICAS:

PRECISIÓN ESTA APUNTA A RECOMENDAR AL USUARIO CONTENIDOS QUE SEAN DE SU INTERESE.

DIVERSIDAD ESTA APUNTA A NO MOSTRAR AL USUARIO CONTENIDOS EXCLUSIVAMENTE DE UN MISMO TIPO

SERENDIPITY ESTE APUNTA A RECOMENDAR AL USUARIO CONTENIDOS QUE NO SON MUY POPULARES EN GENERAL.

ESTE ESTA ASOCIADO AL PRINCIPIO DEL 'LONG TAIL'. ESTE PLANTEA QUE LA DISTRIBUCIÓN DEL CONTENIDO ES QUE HAY POCOS CONTENIDOS POPULARES Y UNA ENORME CANTIDAD DE CONTENIDO CON BAJA POPULARIDAD. ENTONCES EL SISTEMA DEBE APUNTAR SUS SUGERENCIAS A ESTA ENORME CANTIDAD DE PRODUCTOS.

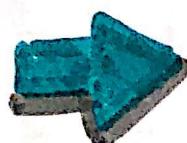
PROBLEMAS TÍPICOS

LA MAYORÍA DE ESTOS PROBLEMAS SON MUY DIFÍCIL DE SOLUCIONAR Y NOS VAN A PERMITIR ENTENDER LA COMPLEJIDAD QUE PODEMOS ENFRENTAR.

- LOS GUSTOS DE LOS USUARIOS PUEDEN CAMBIAR Y NO SABEMOS EL PORQUE.
- LA INFLUENCIA DEL TIEMPO.
- EL USUARIO A VECES QUIERE VER COSAS QUE NO LE GUSTAN
- LO QUE EL USUARIO CALIFICA VS LO QUE EL USUARIO VE.

SISTEMAS NO PERSONALIZADOS

UNA PRIMERA APROXIMACIÓN A SISTEMAS DE RECOMENDACIÓN ES ESTA. CONSISTE EN RECOMENDARLE A TODOS LOS USUARIOS LAS MISMAS COSAS, ES DECIR, SUPONER QUE LA OPINIÓN DE LOS USUARIOS COMO GRUPO ES SUFFICIENTE PARA RECOMENDARLE A CUALQUIERA DE ELLOS. (EJEMPLO: TRIP ADVISOR HACE UN PROMEDIO DE TODAS SUS CALIFICACIONES)



COMO EJEMPLO VAMOS A UTILIZAR **REDDIT**, ACA ES NECESARIO ORDENAR TANTO LOS COMENTARIOS COMO LAS NOTICIAS EN LA PAG PRINCIPAL. EN LOS COMENTARIOS DE LAS NOTICIAS SE PUEDE HACER UPVOTE O DOWNVOTE, PARA INDICAR SI ESTAS A FAVOR O NO RESPECTIVAMENTE.

RECOMENDANDO COMENTARIOS

LAS PRIMERAS DOS IDEAS SON MUY SIMPLES PERO LAMENTABLEMENTE NO FUNCIONAN.

- DIFERENCIA ENTRE UPVOTES Y DOWN VOTES
- TASA DE UPVOTES

INTERVALOS DE CONFIANZA

ESTA ECUACIÓN NOS VA A SOLUCIONAR EL PROBLEMA DE LOS COMENTARIOS QUE TENIAN LOS DOS DE ARRIBA.

ENTONCES DADOS LOS UPVOTES Y DOWN VOTES QUE DEMOS SABEMOS CON UNA CONFIANZA DADA CUÁL ES EL LÍMITE INFERIOR PARA LA PROBABILIDAD DE QUE EL COMENTARIO SEA POSITIVO. ESTE PROBLEMA SE DEDUCE AL CALCULO DEL INTERVALO DE CONFIANZA PARA UN PROCESO DEL TIPO BERNOULLI.

$$\left(\hat{P} + \frac{Z^2}{2N} \pm Z \sqrt{\left[\hat{P}(1-\hat{P}) + \frac{Z^2}{4N} \right] / N} \right)$$

FORMULA DE WILSON

- $\hat{P} = \frac{U}{U+D}$ PROPORCIÓN DE UPVOTES
- $N = U+D$ TOTAL DE LOS VOTOS.
- Z ES EL CUANTIL CORRESPONDIENTE A LA DISTRIBUCIÓN NORMAL DEL % DE LA CONFIANZA

CON ESTA FORMULA PODEMOS CALCULAR LA PROBABILIDAD DE QUE UN COMENTARIO SEA POSITIVO Y ORDENARLOS DE MAYOR A MENOR - ESTO NOS DA UN ORDENAMIENTO INDEPENDIENTEMENTE DEL VOLUMEN DE VOTOS.

ORDENANDO NOTICIAS

LA IDEA ACA ES MOSTRAR LAS NOTICIAS ORDENADAS POR IMPORTANCIA. A TODOS LOS USUARIOS LE DECOMENDAMOS LAS MISMAS. SI BASAMOS EL SCORE DE CADA NOTICIA EN LA MISMA FORMULA QUE USAMOS PARA LOS COMENTARIOS, LAS NOTICIAS NUEVAS SON PENALIZADAS POR NO TENER MUCHOS VOTOS AUN. POR OTRO LADO, A MEDIDA QUE PASA EL TIEMPO QUEDAMOS QUE EL SCORE DISMINUYA, YA QUE PIERDEN RELEVANCIA.

→ SUPONGAMOS QUE UN USUARIO EN REDDIT VISITA EL SITIO CADA UNA CIERTA CANTIDAD DE SEGUNDOS S , LLAMAMOS **TASA DE RELOAD** $\lambda = 1/S$.

TENEMOS CUATRO POSIBLES ESCENARIOS CUANDO UN USUARIO VE UNA HISTORIA.

- ES UNA NOTICIA NUEVA QUE AL USUARIO LE GUSTA **A** NO LE GUSTA **B**
- ES UNA NOTICIA VIEJA QUE AL USUARIO LE GUSTA **C** NO LE GUSTA **D**
- P = PROBABILIDAD DE QUE UNA NOTICIA LE GUSTE A UN USUARIO.
- Q = PROBABILIDAD DE QUE UNA NOTICIA SEA NUEVA

NOTICIA NUEVA Y LE GUSTA NOTICIA NUEVA Y NO LE GUSTA NOTICIA VIEJA Y LE GUSTA

$$U(P,Q) = A * PQ + (1-P)Q * B + P(1-Q) * C + \dots$$
$$(1-P)(1-Q) * D$$

NOTICIA VIEJA Y NO LE GUSTA.

FUNCION DE UTILIDAD (PUNTAJE)

SIENDO LAS PROBABILIDADES

$$\bullet P = \frac{U+1}{U+D+2}$$

$$\bullet Q = e^{-\lambda T}$$

YA QUE $Q \sim \text{POISSON}(1)$

PODEMOS ASUMIR QUE DADA LAS NOTICIAS VIEJAS NO HAY GANANCIA ($C = D = 0$) Y PARA LAS NOTICIAS QUE AL USUARIO LE GUSTAN LA UTILIDAD = 1 ($A = 1$) Y LAS QUE NO LE GUSTAN = -1 ($B = -1$)

ENTONCES (LUEGO DE VARIAS OPERACIONES)

$$U(U, D, A, B) = \log(U-D) + \frac{A-B}{45000}$$

DONDE A ES EL TIMESTAMP DE LA NOTICIA Y B ES EL TIMESTAMP EN QUE FUE CREADO DE REDDIT.

SISTEMAS PERSONALIZADOS

CONTENT BASED

EL OBJETIVO ACA ES RECOMENDARLE AL USUARIO ITEMS SIMILARES A AQUELLOS QUE LE HAN GUSTADO.

POB CADA ITEM ES NECESARIO CONSTRUIR UN PROFILE, ESTE ES UN VECTOR QUE REPRESENTA LAS CARACTERISTICAS DEL ITEM. ES BINARIO 1 SI CONTIENE LA CARACTERICISTA, 0 CASO CONTRARIO. ESTE VECTOR PUEDE TENER MUCHAS DIMENSIONES.

POB CADA USUARIO TAMBIEN CONSTRUIMOS UN PROFILE QUE ES UN VECTOR DE IGUAL DIMENSIONES.

CUANDO UN USUARIO CALIFICA UNA PELICULA LE SUMAMOS EL PUNTAJE A AQUELLAS DIMENSIONES PARA LAS CUALES LA PELICULA TIENE UN '1' EN FEATURE.

TENIENDO EL PROFILE DE CADA ÍTEM Y DE CADA USUARIO PODEMOS ESTIMAR LA CALIFICACIÓN DE UN USUARIO PARA UN ÍTEM CALCULANDO:

$$\cos \theta = \frac{xy}{\|x\| \|y\|}$$

PODEMOS NORMALIZAR (DIVIDIENDO A CADA VECTOR POR SU NORMA) Y ENTONCES EL RESULTADO ESTÁ DENTRO DE 0-1.

VENTAJAS

- SOLO NECESITA LA INFORMACIÓN DEL USUARIO.
- DETECTA LOS GUSTOS PARTICULARES DEL USUARIO.
- PUEDE RECOMENDAR TANTO ÍTEMS NUEVOS COMO VIEJOS.
- PUEDE EXPLICAR EL MOTIVO DE SU RECOMENDACIÓN.

DESVENTAJAS

- ENCONTRAR LOS FEATURES DE LOS ÍTEMS ES MUY DIFÍCIL.
- NO PUEDE RECOMENDAR ÍTEMS FUERA DE LOS GUSTOS DEL USUARIO.
- NO PUEDE RECOMENDAR NADA A LOS USUARIOS NUEVOS.
- NO PUEDE USAR LA OPINIÓN DE OTROS USUARIOS.

➡ ESTE SISTEMA FUNCIONA BIEN PERO ES DIFÍCIL DE CREAR Y MANTENER. NO ES MUY POPULAR.

COLLABORATIVE FILTERING

SEAN N USUARIOS Y M ÍTEMS. LOS USUARIOS CALIFICAN LOS ÍTEMS CON UN NÚMERO DEL 1 AL 5 (PODRÍA SER OTRA MÉTRICA). PODEMOS REPRESENTARLO EN UNA **MATRIZ DE UTILIDAD**

		ÍTEMES				
		SHERLOCK	VELVET GOLDMING	AVENGERS	STAR WARS	THE GODFATHER
USUARIOS	2	2	4	5		
	5	4			1	
		5		2		
	1		5		4	
		4			2	
	4	5	1			

LOS ELEMENTOS QUE FALTAN NO SON REBOS SINO QUE SON NÚMEROS DESCONOCIDOS.

EL OBJETIVO DE ESTE MÉTODO ES ESTIMAR LAS CALIFICACIONES QUE NOS FALTAN EN LA MATRIZ. CF TIENE DOS FORMAS DE ESTIMAR LAS CALIFICACIONES.

1) SEMEJANZA ENTRE USUARIOS

SUPONIENDO QUE QUEDAMOS ESTIMAR LAS CALIFICACIONES QUE LE FALTAN AL USUARIO ' i ', LO QUE HACEMOS ES BUSCAR LOS USUARIOS MÁS PADECIDOS A ' i ' Y CALCULAR LOS FALTANTES EN BASE A UN PROMEDIO PONDERRADO DE LAS CALIFICACIONES DE LOS DEMÁS USUARIOS. PONDERRADAS DE ACUERDO A LA SEMEJANZA QUE TENGA CON NUESTRO USUARIO ' i '.



A MODO DE EJEMPLO VAMOS A ESTIMAR LA CALIFICACIÓN DE SHERLOCK PARA EL USUARIO 5

• BUSCAMOS LOS USUARIOS MÁS SIMILARES.

PARA ESTO NECESITAMOS UNA FUNCIÓN DE SEMEJANZA, QUÉ CALCULE CUAN SIMILARES SON DOS USUARIOS EN BASE A LA CALIFICACIÓN QUE HAN REALIZADOS LOS MISMOS.

$$\text{SIM}(x, y) = \frac{\sum (Q_{xs} - \bar{Q}_x)(Q_{ys} - \bar{Q}_y)}{\sqrt{\sum (Q_{xs} - \bar{Q}_x)^2} \sqrt{\sum (Q_{ys} - \bar{Q}_y)^2}}$$

PERO PODEMOS NOTAR QUE SI LOS PROMEDIOS SON CERO ENTONCES LA FUNCIÓN DE CORRELACIÓN DE PEARSON ES IGUAL AL COSENO DE X, Y.

ENTONCES PARA PODER UTILIZAR EL COSENO (POQUE ES MAS FÁCIL) HAY QUE

- CALCULAR EL PROMEDIO PARA CADA USUARIO, SOLO HAY QUE TENER EN CUENTA LOS COMPLETADOS.
- RESTARLE A CADA CALIFICACIÓN EL PROMEDIO Y ASÍ GENERAR MI NUEVA MATRIZ DE UTILIDAD.
- CALCULAR $\cos \theta = \frac{XY}{\|X\| \|Y\|} \rightarrow$ PRODUCTO INTERNO
PARA SABER LA SEMEJANZA ENTRE LOS USUARIOS

	SHERLOCK	HOUSEM D.C.	AVENGERS	GHOSTBUSTERS	BRAVE	THE
1	2	2	4	5		
2	5		4			1
3			5		2	
4			1		5	4
5					4	
6	4	5	1			

CALCULAMOS
EL PROMEDIO
DE CADA USER

$$M_1 = 3,25$$

$$M_2 = 3,33$$

$$M_3 = 3,5$$

$$M_4 = 3,33$$

$$M_5 = 3$$

$$M_6 = 3,33$$



-1,25		-1,25	0,75	1,15	
1,67		0,67			-2,33
			1,5		-1,5
			-2,33	1,67	0,67
				1	-1
0,67	1,67		-2,33		

SE LO RESTAMOS
A LAS CALIFICACIONES

- $\text{SIM}(5,1) = -0,34$
- $\text{SIM}(5,2) = 0,72$
- $\text{SIM}(5,3) = 0,50$
- $\text{SIM}(5,4) = -0,16$
- $\text{SIM}(5,6) = \text{NA}$

$$\frac{1 * 1,25}{\sqrt{-1,25^2 + -1,25^2 + 0,75^2} * \sqrt{1^2 + 1^2}}$$

NOTAR QUE EN EL PI SOLO TENGO EN CUENTA LAS COORDENADAS QUE TIENEN AMBOS VALORES.

AHORA PODEMOS CALCULAR CADA CALIFICACIÓN FALTANTE:

$$R_{X_i} = \frac{\sum_{y \in N} S_{xy} R_{y_i}}{\sum_{y \in N} S_{xy}}$$

CALIFICACIÓN DEL USUARIO SIMILAR PARA LA PELÍCULA QUE QUIERO.

EN DONDE N ES EL CONJUNTO DE LOS 'N' USUARIOS MAS SIMILARES AL QUE QUEDEMOS ESTIMAR Y R_{y_i} ES EL RATING DEL USUARIO Y PARA EL ÍTEM 'i'. S_{xy} ES LA SEMEJANZA.

ENTONCES PARA EL EJEMPLO BUSCAMOS ESTIMAR EL RATING DE SHERLOCK PARA EL USUARIO 5. USAMOS $N=2$

LOS DOS MAS SIMILARES SON LOS USUARIOS 2 Y 3. PARA LA FORMULA SOLO UTILIZAMOS LOS USUARIOS QUE CALIFICARON AL ITEM QUERIDO.

$$R = \frac{0,72 * 5}{0,72} = 5$$

EL RATING PARA SHERLOCK ES 5 AL USUARIO DEBERÍA GUSTARLE !

NOTAR QUE EL USUARIO 3 NO CALIFICO ENTonces NO SE INCLUYE.

- PARA TOMAR LOS N USUARIOS MAS PARECIDOS AL USUARIO AL QUE QUEDEMOS RECOMENDABLE PODEMOS USAR LSH.

2) SEMEJANZA ENTRE ÍTEMS

AHORA PARA ESTIMAR LA CALIFICACIÓN DE UN USUARIO PARA UN ÍTEM ES BUSCAR LOS ÍTEMS MÁS PARECIDOS AL QUE QUEREMOS ESTIMAR Y QUE HAYAN SIDO CALIFICADOS POR EL USUARIO. LUEGO ESTIMAMOS HACIENDO UN PROMEDIO PONDERADO ENTRE LAS CALIFICACIONES DEL USUARIO PARA ESTOS ÍTEMS.

	users											
	1	2	3	4	5	6	7	8	9	10	11	12
movies	1	1	3			5		5		4		
2			5	4			4		2	1	3	
3	2	4		1	2		3		4	3	5	
4		2	4		5			4		2		
5			4	3	4	2			2	5		
6	1		3		3			2		4		

CALCULAMOS
EL PROM

$$\begin{aligned} \bar{m}_1 &= 3,6 \\ \bar{m}_2 &= 3,16 \\ \bar{m}_3 &= 3 \\ \bar{m}_4 &= 3,4 \\ \bar{m}_5 &= 3,33 \\ \bar{m}_6 &= 2,6 \end{aligned}$$

	users											
	1	2	3	4	5	6	7	8	9	10	11	12
movies	1	1	3			5			5		4	
2			5	4			4			2	1	3
3	2	4		1	2		3		4	3	5	
4		2	4		5			4			2	
5			4	3	4	2				2		5
6	1		3		3			2			4	

LO DESTAMOS (NO LO HICE
POR FALTA DE TIEMPO)

NOTAR QUE EN ESTE CASO LOS ÍTEMS (PELÍCULAS) SON LAS FILAS.



A MODO DE EJEMPLO QUEDAMOS CALCULAR LA CALIFICACIÓN DEL USUARIO 5 PARA LA PELÍCULA 1

- PRIMERO CALCULAMOS LA SEMEJANZA DE TODAS LAS PELÍCULAS CON RESPECTO A LA 1. PARA PODER UTILIZAR EL COSENO, DESTAMOS A CADA CALIFICACIÓN EL PROMEDIO DE LA PELÍCULA

- $\text{SIM}(1,2) = -0,18$
- $\text{SIM}(1,3) = 0,41$
- $\text{SIM}(1,4) = -0,10$
- $\text{SIM}(1,5) = -0,31$
- $\text{SIM}(1,6) = 0,59$

PODEMOS VER QUE LAS MÁS SIMILARES SON LA PELÍCULA 6 Y LA 3.

AHORA, QUE TENEMOS LOS ÍTEMS SIMILARES, CALCULAMOS EL RATING PARA LA PELÍ 1. CON $N=2$.

EN ESTE CASO HACEMOS EL PROMEDIO PONDERRADO DE LAS PELICULAS SIMILARES CALIFICADAS POR EL USUARIO 5.

$$Q = \frac{0,59 * 3 + 0,41 * 2}{0,59 + 0,41} = 2,6$$

NOTAR QUE EL USUARIO 5 PODRÍA NO HABER CALIFICADO ALGUNA DE LAS PELIS SIMILARES

CF: EN BASE A DESVIACIONES

EN ESTE CASO VAMOS A ESTIMAR LA CALIFICACIÓN DEL USUARIO 'i' A LA PELICULA 'j' DE LA SIGUIENTE FORMA.

$$R_{ij} = \bar{m} + \delta_i + \delta_j + \delta_{ij}$$

- \bar{m} = ES EL PROMEDIO DE TODAS LAS CALIFICACIONES DE TODAS LAS PELICULAS
- δ_i = LA DESVIACIÓN DEL USUARIO 'i' CON RESPECTO AL USER. ESTO SE CALCULA COMO LA DIFERENCIA ENTRE EL PROMEDIO DE LAS CALIFICACIONES DEL USUARIO 'i' - EL PROMEDIO GLOBAL (\bar{m}).
- δ_j = LA DESVIACIÓN DEL ITEM 'j' CON RESPECTO A \bar{m} . ESTO SE CALCULA COMO LA DIFERENCIA ENTRE EL PROMEDIO DE LAS CALIFICACIONES DE ESE ÍTEM - EL PROMEDIO GLOBAL (\bar{m}).
- δ_{ij} = LA DESVIACIÓN DEL USUARIO 'i' PARA LA PELICULA 'j'

ESTO SE CALCULA COMO:

RATING DE LA PELÍ
SIMILAR AJ DEL USUARIO i

$$s_{ij} = \frac{\sum_{j \in N} s_{ij} (R_{ij} - B_{xj})}{\sum_{j \in N} s_{ij}}$$

SEMEJANZA

SIENDO $B_{xj} = M + \delta_x + \delta_j^*$

USER ITEM



CALCULAMOS LA DESVIACIÓN DEL USUARIO δ_x CON RESPECTO A LA PELICULA j USANDO LA SEMEJANZA ITEM - ITEM PERO COMPUTANDO EL PROMEDIO PONDERRADO DE LAS DESVIACIONES EN LUGAR DE LAS CALIFICACIONES. RECORDDEMOS QUE N SON LOS N ITEMS MAS SIMILARES.

EJEMPLO

	users											
movies	1	2	3	4	5	6	7	8	9	10	11	12
1	1		3		?	5			5		4	
2			5	4			4			2	1	3
3	2	4		1	2		3		4	3	5	
4		2	4		5			4			2	
5			4	3	4	2				2	5	
6	1		3		3			2			4	

NUEVAMENTE QUEREMOS ESTIMAR LA CALIFICACIÓN DEL USUARIOS PARA LA PELICULA 1.

- CALCULAMOS M
 $111/35 = 3.17 \Rightarrow M = 3.17$

- CALCULAMOS $s_i = M_{USER} - M_{GLOBAL}$
 $M_{USER} = (2+5+4+3)/4 = 3,5$

$$\underline{s_i = 0,33}$$

- DE LA MISMA FORMA CALCULAMOS $s_j^* = M_{ITEM} - M_{GLOBAL}$
 $(1+3+5+5+4)/5 = 3,6$

$$\underline{s_j^* = 0,43}$$

- AHORA CALCULAMOS s_{ij} . EN EL EJEMPLO ANTERIOR YA HABIMOS CALCULADO LA SEMEJANZA A LA PELÍ 1 → PELÍ 3 Y 6

$$s_{ij} = \frac{0,41 * (2 - (3,17 + 0,33 - 0,17)) + 0,59 * (3 - (3,17 + 0,33 - 0,59))}{0,41 + 0,59}$$

DESVIACIÓN
DE LA PELÍ 3

DESVIACIÓN
DE LA PELÍ 6

ESTOS HAY
QUE CALCULAR.

$$\underline{s_{ij} = -0,5}$$

FINALMENTE $3.17 + 0,33 + 0,43 - 0,5 = 3.43$

EVALUACIÓN DE SISTEMAS

EVALUAR UN SISTEMA DE RECOMENDACIÓN ES ALGO COMPLEJO. QUEREMOS SABER SI LO QUE RECOMENDAMOS ES REALMENTE ÚTIL PARA EL USUARIO, LO CUAL ES DE SUBJETIVO. POR ESTA RAZÓN ES DIFÍCIL CONSTRUIR METRICAS QUE SIRVAN PARA EVALUAR Y COMPARAR UN SIST DE RECOMENDACIONES.

BASADAS EN PREDICCIONES

REALIZAMOS PREDICCIONES PARA LOS ITEMS QUE EL USUARIO YA HA CALIFICADO USANDO EL SISTEMA Y LUEGO COMPARAR LOS RATINGS CONOCIDOS CON LAS PREDICCIONES. PARA ESTO USAMOS CROSS VALIDATION (SPÚT DEL 80% Y 20%).

RMSE

$$J = \sqrt{\frac{1}{N} \sum_{ij} (\hat{R}_{ij} - R_{ij})^2}$$

BASADAS EN EL ORDEN

EN ESTA FAMILIA DE METRICAS EL SIST REALIZA LAS PREDICCIONES Y SE ORDENAN DE MAYOR A MENOR, LUEGO COMPARAMOS EL ORDEN DE LOS ITEMS RECOMENDADOS CON EL ORDEN CONOCIDO POR EL USUARIO.

MRR

PARA CADA USUARIO TENEMOS UNA LISTA DE ÍTEMS RELEVANTES, LUEGO DE LAS RECOMENDACIONES BUSCAMOS EN QUÉ NÚMERO DE ORDEN APARECE EL PRIMER ÍTEM Y CALCULAMOS 1/ POSICIÓN.

$$MRR(O, U) = \frac{1}{N} \sum_{U \in U} \frac{1}{P_U}$$

DONDE P_U ES LA POSICIÓN EN LA LISTA DEL PRIMER ÍTEM QUE SABEMOS QUE ES RELEVANTE PARA EL USUARIO.

MAP

UTILIZAMOS PRECISIÓN Y RECALL

$$P = \frac{\text{ITEMS RELEVANTES}}{\text{ITEMS RECOMENDADOS}}$$

$$R = \frac{\text{ITEMS RELEVANTES}}{\text{TOTAL ITEMS RELEVANTES}}$$

CALCULAMOS UN PROMEDIO DE LA PRECISIÓN PARA CADA NIVEL DE RECALL DIFERENTE = AP
Y SI PROMEDIAMOS AP:

$$\text{MAP}(O, U) = \frac{1}{N} \sum_{U \in U} \text{AP}(O(U))$$

CORRELACIÓN DE SPERMAN

EL COEFICIENTE DE SPEARMAN COMPARA LA POSICIÓN O RANKING DE LOS VALORES. ESTO TIENE UN PROBLEMA, CASTIGA TODOS LOS ERRORES DE LA MISMA FORMA. ESTO LO SOLUCIONA NDCG.

NDCG

PRIMERO CALCULAMOS DCG; DIVIDIMOS EL RATING REAL DE CADA ITEM POR \log_2 DE SU POSICIÓN EN LA LISTA DE RECOMENDADOS, PARA LOS DOS PRIMEROS ITEMS DIVIDIMOS POR 1.

$$\text{NDCG} = \frac{\text{DCG}}{\text{PERFECT DCG}}$$

PERFECT DCG → DCG CON LOS RATINGS EN SUS POSICIONES POSTAS

LUEGO PROMEDIAMOS EL NDCG PARA TODOS LOS USUARIOS PARA OBTENER EL PROMEDIO PARA EL SISTEMA.

MODELOS LATENTES

FACTORIZATION MACHINES

LAS CALIFICACIONES QUE LOS USUARIOS REALIZAN DE LOS ÍTEMS APORTAN VALIOSA INFORMACIÓN SOBRE LAS PREFERENCIAS DE LOS MISMOS. ESTE ES UN TIPO DE MODELO LATENTE PORQUE NOS PERMITE INFERRIR PREFERENCIA DE LOS USUARIOS QUÉ NO ESTÁN EXPLÍCITAS EN LA MATRIZ DE UTILIDAD.

ESTE ALGORITMO SE BASA EN ENCONTRAR LOS FACTORES LATENTES ASOCIADOS A CADA VARIABLE DEL SISTEMA. SUPONEMOS QUE TENEMOS { USER-ID, MOVIE-ID, RATING }

CONVERTIMOS USER-ID, MOVIE-ID EN 'N' Y 'M' VARIABLES BINARIAS MEDIANTE **ONE-HOT-ENCODING**.

ESTO NOS LLEVA A UN MODELO DE REGRESIÓN LINEAL

$$\hat{Y} = w_0 + \sum_{i=1}^N w_i * x_i + \sum_{j=1+N}^{M+N} w_j * x_j$$

ESTE MODELO CAPTURA UN COEFICIENTE POR CADA UNA DE LAS VARIABLES DEL MODELO Y ESTIMA LA CALIFICACIÓN EN BASE A ESTOS COEFICIENTES. EN OTRAS PALABRAS LA CALIFICACIÓN VA A DEPENDER DEL USUARIO Y LA PELÍCULA. SE PODRÍA EXTENDER ESTE MODELO AGREGANDO OTROS VARIABLES CODIFICADAS MEDIANTE BHE!

DE TODAS FORMAS, ESTE MODELO NO CAPTURA LA INTERACCIÓN ENTRE LAS DIFERENTES VARIABLES.

AL FACTORIZAR LO QUE HACEMOS ES AGREGAR TODOS LOS COEFICIENTES DE INTERACCIÓN ENTRE CADA UNA DE LAS VARIABLES DEL MODELO, EN TOTAL AGREGAMOS (N_z) NUEVAS VARIABLES DEL TIPO w_{ij} , INDICANDO LA RELACIÓN ENTRE 'i' Y 'j' DEL MODELO

$$\hat{y} = w_0 + \sum_{i=1}^N w_i * x_i + \sum_{i=1}^N \sum_{j=i+1}^N w_{ij} * x_i * x_j$$

ESTO SUGIERE UNA EXPLOSION COMBINATORIA, YA QUE NO HAY FORMA DE MANEJAR TODAS LAS INTERACCIONES POSIBLES.

PARA SOLUCIONAR ESTO FM, DECIDE REPRESENTAR A CADA COLUMNAS (VARIABLES) MEDIANTE UN VECTOR DE k ELEM, SIENDO k LA CANTIDAD DE FACTORES LATENTES QUE QUEDEMOS USAR.

$$\hat{y} = w_0 + \sum_{i=1}^N w_i * x_i + \sum_{i=1}^N \sum_{j=i+1}^N \langle v_i, v_j \rangle * x_i * x_j$$